Accessibility of Multilingual Terminological Resources - Current Problems and Prospects for the Future

Gerhard Budin*, Alan K. Melby#

* University of Vienna
Department of Translation and Interpretation, 1190 Vienna, Austria
<gerhard.budin@univie.ac.at>

# Brigham Young University
Department of Linguistics, Provo, Utah, USA
<akm@byu.edu>

Abstract

In this paper we analyse the various problems in making multilingual terminological resources available to users. Different levels of diversity and incongruence among such resources are discussed. Previous standardization efforts are reviewed. As a solution to the lack of co-ordination and compatibility among an increasing number of 'standard' interchange formats, a higher level of integration is proposed for the purpose of terminology-enabled knowledge sharing. The family of formats currently being developed in the SALT project is presented as a contribution to this solution.

1. Current Problems

Multilingual Terminological Resources (MTRs) have been created for decades for a variety of purposes. The tradition of specialized lexicography or terminography for creating technical dictionaries has developed concomitantly with the rise of modern science and technology; multilingual terminology databases have been in existence since the 1950s.

Being an integral part of the language industries and the information economy, MTRs have been integrated more recently into machine translation systems, technical authoring systems, translation memories and text alignment systems, corpus linguistics applications, controlled language systems, etc. With the needs and requirements of these applications multiplying rapidly, MTRs have diversified even more into different forms, more specifically on the following levels, showing the diversity and incongruence of MTRs:

- *Ontologies*: Domain-specific knowledge organization systems have flourished, generating the need for ontology mapping procedures
- *Categorization*: Terminological information has been categorized in many different ways, calling for standards and mapping procedures
- *Data models*: Relational and object-oriented database management systems allow the definition of a wide variety of data models for MTRs, creating the need for some way to pass information between data models with minimal loss
- *Formats*: Structured information from a given data model can be represented in a variety of XML-based markup systems, each faithful to the same data model and conformant to the rules of XML, yet each with its own DTD/ schema and apparently very different from other formats. This diversity creates the need for a standard format, or at least for interoperability among standard formats.

On all these levels, quite a number of standards have emerged: ISO standards such as ISO 12200 and 12620, industry standards such as LISA's TMX and TBX, project- or organization-specific standards on the international level such as the TEI, and on the EU level such as ELRA, EAGLES, Geneter (from Inesterm and other projects), IIF (from the Interval project), OLIF (from the Otelo project), and many others.

But the integrative effect of standardization has been limited by lack of co-ordination among such initiatives, resulting in the absence of interfaces among application-specific formats and models.

2. The User's Perspective

For a long time, the absence of broadly accepted standards for sharing terminological resources and the incompatibility of competing or overlapping standards has been problematic. Also, the usability of the standards is limited if they cannot be easily implemented in concrete industrial environments. These important criticisms have to be taken into account for any future-oriented activities that aim at improving the situation.

From a user's perspective we might categorize barriers to terminological knowledge sharing as follows:

- *Legal barriers*: MTRs might be available in theory, but in practice owners of such resources are unsure of how to share them with others, e.g. with competitors in their own industrial sector, or with (potential) customers. Copyright and intellectual property issues are basically unsolved in the area of language resources.
- *Economic barriers*: even if legal barriers were removed, MTRs might still be unavailable to certain customers because these resources are frequently prohibitively priced, so that the return on the investment in such resources would come too late in time. Pricing and billing policies and procedures with respect to MTRs have not stabilized.
- *Information barriers*: many potential customers have no information on the existence of such MTRs that might be relevant for them, despite the Internet.

- *Technical barriers*: Owners or potential buyers and users of MTRs might not have the necessary tools to access available MTRs. Some resources are not available on open platforms, but only in proprietary data formats, and conversion tools are rarely available.
- *Methodological barriers*: The methods of preparing MTRs differ quite radically from each other on the level of data modeling, the semantics of data categories used in database design, the method of terminology management chosen, etc.

Although these barriers are all significant, the legal and economic issues are manageable within a multilingual document production chain even in the case of a very large world-wide organization, or when an organization desires to share its terminology freely. Even in such cases, however, the technical and methodological barriers remain because of the diversity and incongruence described above.

Only when one coherent approach is adopted will the sharing of terminological resources increase dramatically.

Industry sectors, such as the language industry, will not compromise in their choice of formats or in their strategies for product development cycles. Information technology companies that produce software and hardware with the need for localizing their products into dozens of languages and with distributed company locations, clearly require a single industry-wide standard that is immediately usable. The localization industry, in particular in its cooperation and coordination platform known as LISA (Localization Industry Standards Association) has started to develop industry standards for terminology resources (TBX), building on the widespread adoption of their XML-based format for exchanging translation memory data (TMX). LISA companies are only willing to accept ‘official’ standards (e.g. from ISO) if these standards directly address expressed goals and live up to clearly formulated requirements. As a liaison organization to the ISO technical committee on terminology (ISO/TC 37), the LISA data-exchange group called OSCAR (Open Standards for Container/content Allowing Re-use) has identified several principles that have to be met by ISO standards in order for them to be considered by industry for implementation. The working group in ISO/TC 37 that is responsible for maintaining existing standards such as ISO 12200 and ISO 12620 and for developing new standards are required to make their standards live up to industry’s requirements.

Another important requirement for interchange standards is that they must support workflow processes in heterogeneous and complex production chains within a company that uses, for example a machine translation system, side-by-side with a translation memory tool, a document management system, a terminology management system, controlled language applications, authoring tools, information and knowledge management systems, spare parts administration tools, digital libraries, and other potential applications. The trend toward integrated IT environments where all such components have to seamlessly interact is the main scenario for the language industry in the future. Without standards that support such interfaces, workflow systems are limited in their effectiveness, since not all language technology components come from the same vendor and are thus not necessarily compatible with each other. Nevertheless, workflow systems are being developed and implemented (Allen 1999, Schubert 1999, see also industrial product environments such as SDL Workflow, and the L&H product strategy that demonstrate this trend). Increased accessibility to MTRs is needed to support these workflow systems.

3. Foundations of Strategic Solutions for Knowledge Sharing

The limited accessibility of MTRs across applications has to be enhanced by various means and on all the levels distinguished above. In the rest of this paper, we deal with overcoming technical and methodological barriers

Most standardization efforts so far have concentrated too much on a specific application context or on a specific linguistic theory or modelling doctrine. By critically reviewing and analyzing these approaches, it seems more fruitful to choose an approach that is focused more on the 'common denominator' across different applications, on mapping diverse ontologies, data models, and categorizations onto each other. The complexity of the task of mapping diverse knowledge organization systems, ontologies, and data category schemes can hardly be over-estimated. Current research activities (e.g. Meo-Evoli & Negrini 1999) have generated methods for switching among classification systems. By establishing inter-operability among various data models and formats, within an overall framework, we can facilitate *horizontal knowledge sharing.*

Knowledge sharing is an essential component of knowledge management. For a successful knowledge management strategy, Davenport & Prusak ask for an adequate ‘Culture of Knowledge Tranfer‘ (Davenport & Prusak, 1998) in order to overcome the problems caused by different cultures, different vocabularies and different frames of reference by creating a common understanding, a common vocabulary and a shared culture among those who want to cooperate under a knowledge management scheme. A *common and clear language with shared meanings* is not only the basis of any particular culture, but also the prerequisite for any successful terminology sharing strategy within and among various organizations.

In order to communicate effectively across cultures and to share MTRs across organizations, we have to apply *meta-standards* (Cox, 1999). Such meta-standards include quality management standards (ISO 9000 family) and basic principles of terminology management such as ISO/TC 37 standards that lay down the principles of terminology management (e.g. how to write a definition, how to coin new terms, how to create a concept system, etc.). The consistent application of such meta-standards is the only possibility to ensure that the methodological barriers mentioned above can actually be overcome in an efficient way.

4. A New Approach in the SALT Project

These ideas of meta-standards and knowledge sharing based on open standards are underlying the SALT project (Standards-based Access to Multilingual Lexical and Terminological Resources). In analogy to the sprawling meta-data initiatives such as the Dublin Core (part of the RDF standard of the World Wide Web Consortium), GILS, (http://www.gils.net/), and ISO/IEC 11179 (for the standardizing and registering of data elements, see http://hmra.hirs.osd.mil/mrc/ for an introduction), a meta-model-based family of formats is now being defined within the SALT project. The SALT approach allows the mapping of many of the existing formats, categorizations, models, ontologies, etc. mentioned above to each other and the transformation of a specific MTR representation into another specific one.

The SALT family of data formats has the following properties:

- It is based on XML, thereby allowing the use of XSL and other XML tools
- It is modular in its structure, i.e. those parts of an ontology or elements of lexico-terminological information that are actually relevant for a specific target application can be selected and processed by transformation tools.
- A freeware toolkit will be available on the SALT server (http://www.loria.fr/projets/SALT/) in the year 2001
- It is internationalized, i.e. fully UNICODE enabled.
- It is end-user oriented, distinguishing different user groups of equal importance, industrial tools developers, service providers, translators, technical writers, localizers, and other 'real' end-users.

On January 1st, 2000, the two-year SALT project started in the framework of the EU, funded by the EU Commission as a project in the HLT (Human Language Technologies) sector of the IST (Information Society Technologies) Programme (5th Framework Programme).

The remainder of this paper consists of a brief technical introduction to the SALT family of data models and formats and an explanation of how various groups, both public and private, are co-operating with the SALT project in order to avoid a proliferation of incompatible standards for accessing MTRs.

5. A Technical Introduction to the SALT Family of Data Models and Formats

A data model and, consequently, an XML representation format for a data model must include three logical components: (a) a set of the data element types that are allowed in the model, (b) the permissible content of each data element type, which may be a data type (for example, ISO date) or a list of permissible values for each data element type, and (c) the structural relationships that are allowed among the data element instances. The third component defines the form and the first and second components define the content of the representation.

A basic assumption of the SALT project is that no single data model can possibly serve the needs of all groups who access MTRs. Typically, no format would make use of all the data categories (i.e., data element types) in ISO 12620, which is intended to be an exhaustive inventory. When a representation format is processsed, for example, when an exchange file is imported into an application, each data category allowed in the format must be accounted for, including its permissible content. Therefore, user groups are inclined to disallow unneeded data categories from their data model.

One way to accommodate the needs of various user groups is to define one complex all-inclusive master format that contains all possible data categories and their values and then to define subsets of that monolithic format. One difficulty with such an approach is that such a master format is necessarily unstable. That is, as each new data category is allowed, the format must change to allow that new data category. The master format must even be modified to allow for one new permissible value for one data category among hundreds. Thus maintenance of such a format becomes a nightmare. Or, on the other side of the coin, if the format is frozen and not allowed to change, in which case industry will quickly abandon it.

Another difficulty with the monolithic approach involves writing flexible routines to process an instance of the master format or any subset thereof. Although general-purpose XML parsers can be embedded into end-user applications, error-messages from general-purpose parsers must be contextualized in order to be helpful to a non-expert. This means that the application must understand the XML DTD or schema of the format. The more complex the DTD or schema, the more *expensive* it is for an application to understand it sufficiently well to present a friendly user interface.

An obvious solution to these problems of maintenance and friendliness is the time-tested approach of separating form and content. An example from the history of syntactic theory is to compare the original 1957 version of Generative Grammar, in which all the rules, including the lexical rules, were in one monolithic list, with later versions of generative grammar in which the lexicon has been split off as a module separate from the structural module.

The SALT approach separates form and content in a fashion consistent with an international standard for defining terminological formats (ISO 12200). ISO 12200 and ISO 12620 are the form and content components of a family of formats for representing MTRs. ISO 12200 does not define a particular format; instead it defines a family of formats by showing the structural relationships between meta-data-categories, such as *descriptive element* and *administrative element*, rather than specific data categories, such as *definition*, *contextual example*, or *modification date*. Thus, the structure defined in ISO 12200 even though it must be amended from time to time, is immune to minor changes in data categories and therefore much more stable than the DTD/schema of a monolithic format. Arriving at a content specification for a particular user group may require considerable advance negotiation, as indicated in the title of ISO 12200. The structure of ISO 12200 combines with a particular negotiated content specification to define a particular format.

Current projects within ISO Technical Committee 37 are aimed at (a) defining a very high-level meta-model that leaves room for both XML-based representation formats and relational database design, (b) defining specific XML formats, such as those found in the MSC family (an application of ISO 12200 and 12620), within the broad possibilities allowed for by the meta-model, and (c) providing for interoperability between specific formats, that is, for bi-directional conversions between formats with little or no loss of information, so long as the content specification is held constant. Obviously, if one format makes a distinction between definitions and contextual examples while another format does not, then that distinction will be lost when terminological information is passed through the less nuanced format. No amount of structural manipulation can compensate for incommensurate sets of data categories.

The SALT project is adopting the ISO approach just described and adding to it elements for representing information from machine translation lexicons and other NLP resources. Furthermore, the SALT project recognizes the need for an approach to designing relational databases that corresponds directly to the meta-model approach to defining XML-based representation formats. Granted, these days XML representations are being used more often as a direct basis for query and processing without passing through a relational database, and thus the distinction between representation format and processing format is being blurred. However, this simply emphasizes the need for parallel XML and relational database methodologies for MTRs. One such approach to designing relational databases for MTRs, called Reltef™ is freely available (see http://www.ttt.org/clsframe/reltef.html) and has been implemented to support central terminological databases in a multinational medical technology company, a university project in Spain, and the United Nations offices in Vienna. The Reltef approach should easily be adapted to object-oriented or hybrid databases.

The integrative picture of the SALT project that emerges from the inclusion of NLP lexicons, relational databases, and a meta-model can be outlined as follows:

- At the highest level, the meta-model level, the abstract structure of MTRs is represented using an application-independent diagramming method such as ORM (Embley et al 1992). The meta-data-categories at this level are treated as object classes, and the structural aspect of the meta-model shows relationships between object classes. (See Figure 1: The Meta-model). The structure of a data category specification is also given at this level, using some meta-data formalism such as RDF (a World Wide Web standard), but no particular set of data categories is given except the master inventory in ISO 12620.
- At the intermediate level, the conceptual data-model level, a split occurs reflecting whether the MTR is represented in XML or in a database. Since the emphasis of this paper is the sharing of MTRs, we will discuss the definition of XML data models. All data models are based on the same core structure, which is compatible with the abstract structure in the meta-model. The core structure is expressed as an XML DTD or schema that is compatible with ISO 12200 as amended. Each data model is defined by the logical combination of the core structure and a particular data category specification (DCS). At the middle level, a DCS is expressed as an instance of an XML schema that uses tag names that are intuitive to a terminologist while being equivalent to the RDF specification structure defined at the meta-model level. (See Figure 2: Example of a DCS file).
- At the lowest level, the specific data-model/format level, conceptual data models defined at the intermediate level are instantiated as actual data models implemented in database management systtems or as actual XML formats. One conceptual data model from the intermediate level can have several interoperable formats associated with it at the lowest level. For example, one format may be very similar to the core structure and thus use meta-data-category-like tag names that are specialized by the value of a *type* attribute while another format may have many more specific tag names and be very similar to the Geneter format, which is one particular format for one particular conceptual data model of the SALT family. (See Figure 3: Comparison between Geneter and MSC). One important benefit of the SALT approach is that various Geneter subsets can be generated automatically by a terminologist who has access to the SALT toolkit but who does not know how to write or modifiy an XML schema.

We will call the meta-model level (Level 1), the conceptual data-model level (Level 2), and the data-model/format level (Level 3). (See Table 1: A Table of Levels). Those familiar with the firstness / secondness / thirdness distinction of the philosopher C.S. Peirce, might notice the following analogy (Peirce 1991). Level 1, the meta-model, is connected with firstness in that it represents the potential for many formats but specifies no particular one of them. Level 2, the conceptual data-model level, is the level most closely tied to seconondness in that a particular data category specification, which is the major contribuiton of level 2, is an expression of the requirements of a particular real-world user. Level 3, the data-model/format level, is connected to thirdness in that a particular data model or format is a set of rules for representing. Those rules are abstracted away from the particular user needs that suggested it and can be applied to new situations and sets of data.

The various formats and databases that are implementations of a particular data model are all guaranteed to be interoperable, unlike arbitrary subsets of a monolithic format, and all data models have the same core structure. Thus, even distinct data models are interoperable up to the limits of the ability to map between the data categories and data-category values in their respective specifications. This interoperability is coupled with diversity to overcome the incongruence that has plagued access to MTRs until now.

6. Prospects for the Future through co-operation and co-ordination

The key to the success of the SALT project, which is defining, testing, and refining the format-family approach just described, is co-operation, resulting in co-ordination between the various groups involved. The key groups besides the SALT team proper are (1) ISO Technical Committee 37 (TC37), (2) the LISA OSCAR group (OSCAR), (3) newly formed OLIF2 consortium that brings back together former participants in the Otelo project (Thurmair & Ritzke & McCormick 1998), (4) developers of language technology, and (5) maintainers of proprietary terminological databases.

Although end users of MTRs will be the principal beneficiaries of successful efforts to improve accessibility, they are powerless to effect the needed changes, except by demanding system developers and maintainers implement standards.

The key to the implementation of the SALT standards by the developers of language technology is acceptance by OSCAR, since the central language technology tools for MTRs involve translation-oriented technology and most the major developers of translation tools (including Star, Trados, SDL, and Logos) are represented on the OSCAR Steering Committee (source: March 2000 LISA Forum presentation on OSCAR) and thus have a voice in the development of OSCAR standards. A sufficient number of language technology developers participate in OSCAR to ensure that OSCAR adoption of a standard will spread throughout the commercial developer community. Members of the SALT project have been involved in OSCAR from its beginning. Likewise, OSCAR is an official Category A liaison organization to ISO/TC37 and has strongly influenced the development of ISO 12200 and current projects. Coordination with maintainers of proprietary terminological databases is harder to achieve, since there is no organization for them that is parallel to OSCAR. However, inclusion of Geneter, which has been used in several EU projects, into the SALT family provides some degree of coordination with proprietary databases. To complete the web of needed co-ordination, a representative of the OLIF2 consortium was just recently voted in as a member of the OSCAR Steering Committee, and meetings are scheduled between the SALT project and the technical director of the OLIF2 consortium

Finally, after years of work that seemed to be headed in the direction of multiple incompatible standards, the future of accessibility for Multilingual Terminological Resources looks bright.

7. References

Allen, J., 1999. Adapting the Concept of 'Translation Memory' to 'Authoring Memory' for a Controlled Language Writing Environment. In *Proceedings of the Twenty-first International Conference on Translating and the Computer 10-11 November 1999, London* (unnumbered). London: ASLIB IMI

Budin, G. et al., 1999. Integrating Translation Technologies Using SALT. In *Proceedings of the Twenty-first International Conference on Translating and the Computer 10-11 November 1999, London* (unnumbered). London: ASLIB IMI

Cox, C., 1999. Meta Standards: Tools for Harmony Within Cultural Diversity. In *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering 23-27 August 1999, Innsbruck, Austria* (694-700). Vienna: TermNet

Davenport, T.H. & Prusak, L., 1998. *Working Knowledge. How Organizations Manage What They Know*. Boston, Mass.: Harvard Business School Press

Embley, D. Kurtz. B., and Woodfield, S., 1992. *Object-oriented Systems Analysis: a Model-driven Approach*. New Jersey: Prentice Hall

ISO 12 200: 1999 *Computer-Applications in Terminology – Machine-readable Terminology Interchange Format (MARTIF) – Negotiated Interchange*. Geneva: International Organization for Standardization

ISO 12 620: 1999 *Computer Applications in Terminology – Data Categories*. Geneva: International Organization for Standardization

Melby, A., 1998. Data Exchange Standards from the OSCAR and MARTIF projects. In *Proceedings of the First International Conference on Language Resources and Evaluation. 28-30 May 1998, Granada, Spain,* (Vol. 1, 3-8). Edited by Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada. Paris: ELRA

Melby, A. & Wright, S.E., 1999. Leveraging Terminological Data for Use in Conjunction with Lexicographical Resources. In *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering 23-27 August 1999, Innsbruck, Austria* (544-569). Vienna: TermNet

Meo-Evoli, L. & Negrini, G., 1999. CoReC: A Model for Integrating Classification Systems. In *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering 23-27 August 1999, Innsbruck, Austria* (293-306). Vienna: TermNet

Murray-Rust, P. & West, L., 1999. Terminology, Language, Knowledge on the Web: Some Advances Made by VHG™ In *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering 23-27 August 1999, Innsbruck, Austria* (618-624). Vienna: TermNet

Peirce, C.S., 1991. *Peirce on Signs. Writings on Semiotic by Charles Saunders Peirce. Edited by James Hoopes*. (184-189). Chapel Hill and London: The University of North Carolina Press

Schmitz, K-D., 1998. ISO 12200 (MARTIF). Basic Concepts and Testing Report. In: *Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium; Tools for Multilingual Communication/TAMA '98* (303-316). Vienna: TermNet

Schubert, K., 1999. Resource and Workflow Management Support in Teletranslation. In *Proceedings of the Twenty-first International Conference on Translating and the Computer 10-11 November 1999, London* (unnumbered). London: ASLIB IMI

Thurmair, G. & Ritzke, J. & McCormick, S., 1999. The Open Lexicon Interchange Format (OLIF). In *Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium; Tools for Multilingual Communication/TAMA '98* (237-262). Vienna: TermNet

Acknowledgements:

We are grateful to valuable comments on the draft paper by Sue Ellen Wright and Klaus Dirk Schmitz. The SALT project is funded under the HLT area of the IST programme in the 5th Framework Programme of the EU.

Index of Acronyms:

DCS: Data Category Specification

DTD: Document Type Definition

EAGLES: Expert Advisory Group on Language Engineering Standards

ELRA: European Language Resource Association

E-R Diagram: Entity-Relationship Diagram

GILS:Government Information Locator Service

IIF: Interval Interchange Format

ISO: International Standards Organisation
LISA: Localization Industry Standards Association
MARTIF: Machine-Readable Terminology Interchange Format
MSC: MRTIF with Specified Constraints
MTR: Multilingual Terminological Resource
NLP: Natural Language Processing
OLIF: Open Lexicon Interchange Format
ORM: Object Relationship Modeling
OSCAR: Open Standards for Container/content Allowing Re-use
OTELO: Open Translation Environment for Localization
RDF: Resource Description Framework
SALT: Standards-based Access to Multilingual Lexicons and Terminologies
SGML: Standard Generalized Markup Language
TBX: TermBase eXchange format
TEI: Text Encoding Initiative
TMX: Translation Memory eXchange format
UML: Universal Modelling Language
XML: eXtensible Markup Language
IT: Information Technology

Figures:

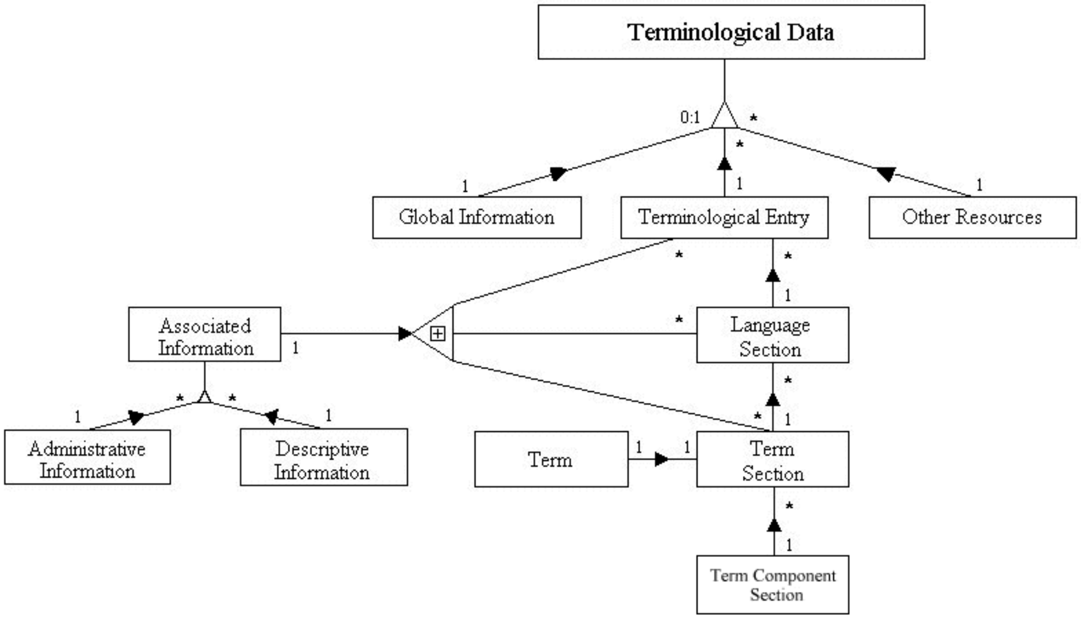

Figure 1: The Meta-model

```
<?xml version="1.0"?>
<martifDCS name='MSCd-supplier' version="0.3" lang='en' xmlns="x-schema:MTFssV03.xml">
<header><title>Supplier Example</title></header>
<datCatSet>
<termNoteSpec name="termType" position="2.1.x">  <!--  position is location in ISO 1260 -->
   <contents datatype="picklist"   targetType="none"> internationalism fullForm partNumber </contents>
</termNoteSpec>
<descripSpec name="subjectField" position="4">
   <contents datatype="picklist"  targetType="none">manufacturing finance</contents>
   <levels>termEntry</levels>
</descripSpec>
<descripSpec name="definition" position="5.1">
   <contents datatype="noteText"/>
   <levels>termEntry langSet</levels>
</descripSpec>
</datCatSet>
</martifDCS>
```

Figure 2: Example of a DCS file

1: <?xml version="1.0" ?>	1: <?xml version="1.0"?>
2: <!DOCTYPE geneter PUBLIC "ISO 12200:1999//DTD GENETER term-only//EN">	2: <!-- use schema so hide doctype "ISO 12200:1999//DTD MARTIF core (MSCcdV03)//EN" -->
3: <geneter profile='ora-gtr' character-set='utf-8'>	3: <martif type="MSC-SRa" lang="en" xmlns="x-schema:MSCga-supplierV03.xml"> <martifHeader> <fileDesc><sourceDesc><p>from an Oracle termbase</p></sourceDesc></fileDesc> <encodingDesc><p type="DCSName">MSCd-supplierV03</p></encodingDesc> </martifHeader> <text><body>
4: <terminological-entry>	4: <termEntry id="ID67">
5: <lil>	
6: <lil-admin-g><entry-identifier>67</entry-identifier></lil-admin-g>	
7: <lil-descrip-g>	
8: <subject-field>manufacturing</subject-field>	8: <subjectField metaType="descrip" value="manufacturing"/>
9: <definition working-language='w-en'>A value between 0 and 1 used in...</definition>	9: <definition metaType="descrip">A value between 0 and 1 used in...</definition>
10: </lil-descrip-g>	
11: </lil>	
12: <ldl language='en'>	12: <langSet lang="en">
13: <tl form-type='full-form'><term>alpha smoothing factor</term></tl>	13a: <tig> 13b: <term>alpha smoothing factor</term> 13c: <termType metaType="termNote" value="fullForm"/> 13d: </tig>
14: </ldl>	14: </langSet>
15: <ldl language='hu'>	15: <langSet lang="hu">
16: <tl><term>Alfa sim...</term></tl>	16: <tig><term>Alfa sim...</term></tig>
17: </ldl>	17: </langSet>
18: </terminological-entry>	18: </termEntry> </body></text>
19: </geneter>	19: </martif>

Geneter-MSC comparison:

Line 1: Both Geneter and MSC are XML applications, so they both begin with an XML processing instruction.

Line 2: Geneter is oriented toward SGML DTDs, while MSC is oriented toward XML schemas. Nevertheless, MSC does have a DTD. It is simply commented out in this example.

Line 3: A Geneter file is an instance of the <geneter> element, an MSC file is an instance of a <martif> element. A subset of Geneter is named using the profile attribute, while a subset of MSC is specified in the encoding description element of the Martif header.

Line 4: A terminogical entry in the meta-model is called a terminological-entry element in Geneter and called a termEntry element in MSC, but they are equivalent.

Lines 4-18: The associated information at the terminological entry level goes inside the lil element in Geneter, while it goes before the first language section in MSC without an explicit element name. Although the format is different in Geneter and MSC, the rest of the information in the entry is basically the same in Geneter and MSC: an entry identifier (67), a subject field (manufacturing), a defintion, an English term that is a full form and a Hungarian term. The language of the definition in Geneter is explicitly indicated by a working-language attribute, while in MSC the lang attribute on the martif element indicates that the content of all elements in this file is English until otherwise specified or inherited.

Line 19: The Geneter entry ends with a </geneter> tag while the MSC entry ends with a </martif> tag.

Clearly, although these two formats differ in a number of details, they are basically equivalent, and it should be possible to automatically convert one to the other.

Figure 3: Comparison between Geneter and MSC

Level	**Components**		**Specification method**	
Level 1: Meta-model	➢ Structural meta-model ➢ Content meta-model ➢ ISO 12620 and how to specify a subset of it		ORM (or UML) RDF (and/or ISO 11179)	
Level 2: Conceptual data model	Relational Database Management ➢ Reltef	XML representations ➢ XLT family ➢ MSC family ➢ Core structure and specific sets of data categories, each defines a conceptual data model	E-R diagram	DTD or an XML schema and a DCS file (each is an XML document)
Level 3: Specific data model/format	Specific relational data models, e.g. Medtronic, United Nations Office in Vienna, University of Granada	Specific XML formats, e.g. each particular subset of MSC primary and secondary representations or Geneter and subsets, or TBX (an XLT subset)	Specific E-R diagrams specifying specific data models in ORACLE, SQL server, Ingres, etc.	Specific DTD or XML schema and DCS file, e.g. Geneter DTD

Table 1: A Table of Levels